FAKULTA APLIKOVANÝCH VĚD ZÁPADOČESKÉ UNIVERZITY V PLZNI
STUDENTSKÁ VĚDECKÁ KONFERENCE 2014

Online Speaker Diarization

Marie Kunešová[1]

1 Introduction

In automatic speech processing, speaker diarization is the task of distinguishing between different speakers within an audio recording and identifying the intervals in which they are active, or in other words, determining "Who spoke when?". This is generally done without any prior knowledge about the actual identities and number of speakers.

The information obtained from speaker diarization can be used in several areas, such as audio indexing and searching or for improving the performance of speech recognition systems. Some of these areas require the diarization to be done online. This represents a more difficult variant of the task and generally leads to a worsened performance compared to offline systems.

An online diarization system was created based on the one proposed by Markov and Nakamura (2007). The goal was to further improve its performance.

2 The Diarization System

The online diarization system uses Gaussian Mixture Models (GMMs) to represent the individual speakers. The basic principle is as follows:

The system starts with only two GMMs, one for each gender, which are trained in advance. The audio stream is divided into short segments and for each of them, the system decides if the segment corresponds to an already known speaker or a new one by comparing the likelihoods of the gender dependent and speaker models. In the case of a new speaker, a new model is created by copying one of the gender dependent models. Otherwise, one of the existing models is selected. The assigned model is then adapted using the data from the segment.

2.1 Speaker Clustering

One of the most problematic areas of the system is the selection of the decision threshold which is used to decide whether a speech segment belongs to a new or old speaker. If this threshold is set too low, multiple speakers will be assigned the same model. Conversely, if it is too high, multiple models will be assigned to the same speaker. Both situations can dramatically reduce the overall performance, yet it is generally impossible to eliminate them both.

For this issue, the following solution was chosen: to select a higher decision threshold, but implement an additional clustering algorithm that would identify any models that are likely to correspond to the same speaker and rectify the situation. For this purpose, a model distance measure based on the cross-likelihood ratio (CLR, Reynolds et al. (1998)) is used.

Once the system decides that two of the models represent the same speaker, there are several possible courses of action. The following two approaches have been examined:

[1] student of the doctoral study programme Applied Sciences and Computer Engineering, field Cybernetics, e-mail: mkunes@kky.zcu.cz

1. Transforming the two models into one
   At the moment, this is done by replacing the models by a single new GMM trained on all the data originally assigned to both. Eventually, a simpler method will have to be found, as this causes a great delay, but this represents an "ideal" resulting model.

2. Keeping both models but considering them to be the same speaker
   Any time one of the models is assigned a new segment, both are updated. After several updates, they will be almost identical. At that point, one of them can be safely removed.

Additionally, offline clustering was performed as well, to serve as a baseline.

2.2 Results

Experiments were done on a set of recordings from Czech parliament meetings with a total of 30 hours of labelled audio. For performance evaluation, the Diarization Error Rate (DER) was used, as described by NIST (2009). It is defined as the fraction of time that is not correctly assigned to a speaker or to non-speech. A forgiveness collar of 0.25s around the reference speaker boundaries was used. For the "without clustering" and "updating both models" variants, the system latency was 2s.

Table 1 shows the achieved results. The *online* column represents the immediate decisions and *final* refers to the results after everything has been retroactively relabelled.

	online	final
Without clustering	9.13	—
Offline clustering	—	5.48
GMM retraining	8.04	5.98
Updating both models	7.75	5.78

Table 1: Comparison of the diarization performance in terms of DER (%)

3 Conclusion

With the use of speaker clustering, the performance of an online speaker diarization system has been significantly improved, with the better of the two compared approaches having a 15% decrease of DER for the immediate decisions and a 37% decrease for the final results, approaching the value obtained from offline clustering.

References

Markov, K., and Nakamura, S., 2007. Never-Ending Learning System for On-line Speaker Diarization. *IEEE Workshop on Automatic Speech Recognition & Understanding, 2007. ASRU.* pp 699–704.

Reynolds, D., Singer, E., Carlson, B., O'Leary J., McLaughlin, J., and Zissman, M., Blind clustering of speech utterances based on speaker and language characteristics, *Proceedings, International Conference on Speech and Language Processing*, vol. 7, pp. 3193–3196.

The NIST Rich Transcription 2009 (RT'09) Evaluation, NIST, 2009 [Online]. Available: http://www.itl.nist.gov/iaui/894.01/tests/rt/2009/docs/rt09-meeting-eval-plan-v2.pdf